结合心智理论的计算语言习得

Andy Liu美国加利福尼亚州克莱蒙特哈维穆德学院{ajliu}@g.hmc.edu

Hao Zhu、Emmy Liu、Yonatan Bisk、Graham NeubigL Technologies Insti 语言卡内基梅隆大学，美国宾夕法尼亚州匹兹堡{zhuhao, mengyan3, ybisk, gneubig}@cs.}cmu.edu

arXiv:2303.01502v1 [cs.CL] 2 Mar 2023

摘要

与当前最先进的语言模型不同，幼儿会通过与周围环境和照护者互动来主动习得语言。有人认为，在语言学习中，推断社会环境中其他智能体心理状态的能力至关重要；Premack 和 Woodruff（1978）将其称为心智理论（ToM）。受 Rabinowitz 等（2018）和 Zhu 等（2021）提出的现代操作化心智理论方法启发，我们构建了具备心智理论的语言学习智能体，并测量其对学习过程的影响。我们通过为说话者智能体赋予一个内部听者模型来模拟心智理论；该模型与说话者一同训练，并用于对备选话语重新排序。我们通过改变任务难度进行实验，并假设模型会习得更复杂的语言，以适应更强的环境压力。我们发现，在图像指称游戏设置中，对心智理论听者组件赋予较高权重来训练说话者，能够提升其表现。我们还发现了一些证据，表明提高训练过程中的任务难度，会使模型在评估中产生更流畅、更精确的话语。这表明，在语言习得的计算模型中进一步融入心智理论以及儿童语言习得的其他见解，可能具有实用价值[1]。

1 引言

人类语言从根本上受到具身世界中的社会交际目标塑造。发展心理学的现代理论通常认为，人类能够快速习得并调整语言，是因为人类具备将心理状态归因于其他智能体的能力（Tomasello，2005）；这种能力也被称为心智理论（ToM）。

此前已有一些研究尝试对心智理论进行计算建模。例如，研究表明，类似心智理论的机制能够帮助模型更好地预测未来智能体的行为（Rabinowitz 等，2018）、在协商（Cao 等，2018）或合作博弈（Bard 等，2020）中建模智能体的信念，或根据听者的语言能力选择恰当的话语（Zhu 等，2021）。然而，心智理论在计算语言习得这一更高层次语境中的影响尚未得到研究。

本文研究了内部心智理论机制和外部环境压力如何共同促进语言学习。我们采用图像指称游戏设置，由说话者和听者之间的一系列训练回合组成，其中说话者代表语言学习者（Zhu 等，2022），听者代表一名流利的教师。当面对一组图像（其中一幅为目标指称对象）时，说话者必须学会生成一段英语话语，使听者能够据此选出目标。若说话者生成的话语被用于正确猜出目标图像，就会获得奖励。此外，还可以根据听者对所选图像的信心程度向说话者提供反馈。这一设置为检验各种奖励信号或模型设计对说话者所习得语言的影响提供了理想的测试平台；语言习得中的语用学研究（例如 Andreas & Klein (2016)）也采用过类似设置。

[1]代码和数据可在 https://github.com/neulab/ToM-Language-Acquisition 找到。

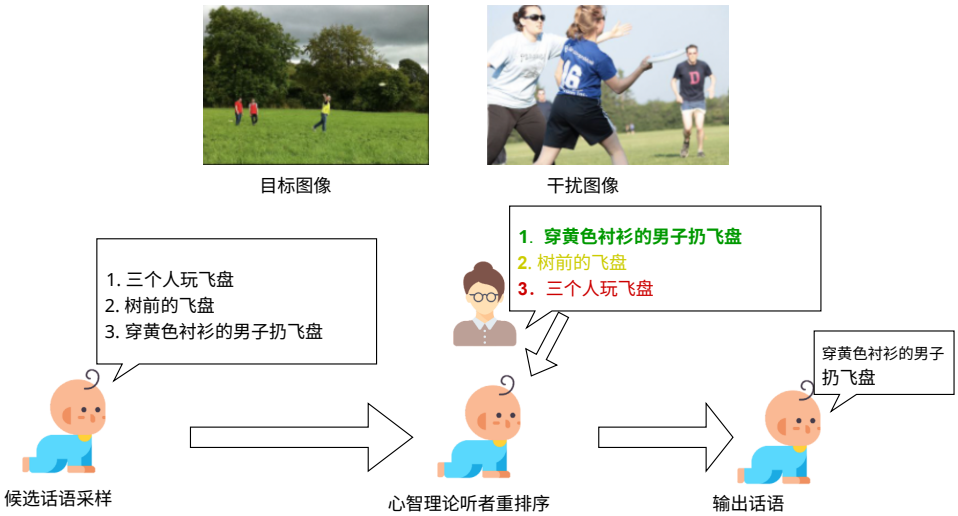

图 1：在我们的实现中，说话者模型使用心智理论的示例。说话者首先生成候选话语，然后根据内部心智理论听者为每个目标图像—话语对给出的概率，对这些候选话语进行重排序。随后，它从候选池中选择得分最高的话语或随机选择一个话语作为输出。

在这一设置下，我们希望更好地理解具备心智理论和不具备心智理论的模型如何适应其环境。我们聚焦于该领域的两个具体研究问题：

RQ1. 在语言习得说话者模型中纳入心智理论，会如何影响其性能和所习得的语言？（内部心智理论机制）RQ2. 在语言习得过程中，我们的模型如何适应更困难的指称游戏环境？（外部环境压力）

我们通过在说话者内部建模一个内部听者模块来研究心智理论的影响（RQ1）。该模块旨在预测哪些话语最有可能带来期望的听者行为。如图 1 所示，我们将心智理论听者为目标图像给出的概率纳入话语重排序过程，从说话者的概率分布中采样并选择更具语用性的话语。为了研究环境压力（RQ2）对说话者的影响，我们通过从不同分布中采样干扰项，创建具有不同难度的指称游戏。这些分布基于各种图像表征模型计算出的图像相似度，包括 CLIP（Radford 等人，2021）、RoBERTa（Liu 等人，2020）和 TF-IDF 变体。

在实验中，我们发现：（RQ1）包含心智理论组件的说话者模型，在流畅度和最终准确率方面，通常优于不包含这些组件的模型。我们还发现：（RQ2）使用视觉和语义上更相似的干扰指称对象进行训练，会使说话者模型生成更长、更流畅且更精确的话语，以区分潜在的指称对象，尽管这并不总能提升指称游戏表现。这些结果表明，在这一设置下，心智理论和环境压力都有助于语言习得。我们仍然发现，我们的说话者模型所习得的语言与人工描述之间存在显著差距。此外，我们还限制了说话者的词汇量和话语最大长度。这两点都表明，这类模型仍有改进空间。不过，我们希望这些结果仍能为语用语言模型提供更好的训练方法，并加深对人类语言习得的计算理解。

2 图像指称游戏环境

沿用 Lazaridou 等人（2016）；Lowe 等人（2019）；Zhu 等（2022）的做法，我们考虑基于真实世界图像的图像指称博弈：说话者和听者合作，从候选图像集合 C 中随机抽取的干扰项里识别目标图像 $x \in C$。只有说话者知道目标图像的身份。说话者生成一段英语话语 u

来描述目标图像，然后将其传递给听者。如果听者理解该话语的概率足够高，就会选择一幅图像 $\hat{\mathbf{x}}$；如果不理解，则会拒绝采取行动。听者还可以决定是否向说话者提供反馈（以目标图像的真实描述形式呈现的语言输入）。

说话者训练由奖励函数驱动，该函数旨在对交际目标进行建模，即引导听者找到正确的指称对象。如果听者选择了目标图像，说话者将获得 1 的正向奖励；如果听者选择了错误的目标图像，则会受到 -1 的惩罚。此外，如果听者没有选择任何图像，说话者还会受到较小的惩罚 $w_{\texttt{noop}}$；这样做是为了惩罚说话者生成含糊不清的话语。

2.1 智能体形式化

我们采用 Zhu 等人（2022）定义的说话者和听者形式。说话者是一个模型$f\colon\ I \to \Sigma^*$，它生成与给定图像相对应的话语（由其词汇中的词元组成的序列）。我们将 Σ^* 定义为说话者词汇，将 I 定义为候选图像集合。

听者定义在 Σ^*和 I^N上，其中 I^N是从 I中采样得到的一组 N候选图像。给定话语$\in\ \Sigma^*$和观测空间 $\in\ I^N$，听者将返回其预测图像的索引；如果无法预测图像，则返回 noop。它也可以返回其词汇中的真实描述。我们用模型 $g\colon\ \Sigma^* \times I^N \to ([0, 1, \ldots N-1] \cup \text{noop}) \times \Sigma^*$来定义它。

2.2 说话者设计

我们使用的说话者是一个图像描述模型，由一个预训练 ResNet（He 等，2016）模型和一个基于 LSTM 的话语生成模型组成。说话者使用 ResNet 模型生成目标图像的嵌入向量后，利用 LSTM 网络自回归地生成话语$u = \{u_i\}_{i=1}^{M}$：

$$\begin{aligned} &P(u_i \mid u_1, u_2, \ldots, u_{i-1}, x) \\ &\propto \exp(w_{u_i}^T \texttt{LSTM}(w_{u_1}, w_{u_2}, \ldots, w_{u_{i-1}}, h_0 = \texttt{ResNet}(x))), \end{aligned} \tag{1}$$

其中，$w_{u_i} \in \mathbb{R}^{d_w}$ 是 u_i的词嵌入。说话者会生成一系列词元，直到达到最大长度（在我们的实验中设为 20），或生成序列结束词元。在我们的实验中，我们使用大小为 200 的词汇，以限制动作空间的规模，并更真实地模拟语言学习者的词汇。

2.3 听者设计

给定一组候选图像 I_j（$j \in 1,\ 2 \ldots n+1$），其中 n是干扰项的数量，以及一个话语 u，听者会为每个图像计算嵌入向量 L（I_j），并为该话语计算嵌入向量 L（u）。随后，它计算每个 L（I_j）与 L（u）的点积，并对这些点积应用 softmax，以计算概率 P（$I_j|u$）$\propto \exp$（L（I_j）L（u））$\cdot$，表示该话语指代每个图像 I_j 的概率。

听者还包含一个基于规则的组件，用于控制何时以真实描述的形式提供语言输入 $C(x)$。我们引入两个阈值，θ_1 和 θ_2。听者计算出最可能的目标图像 $t = \arg\max\ P(I_j|u)$，以及赋予该选择的概率 $P_{\max} = \max\ P(I_j|u)$ 后，会根据以下逻辑返回其选择 $\hat{x}$ 并提供输入 $f_{\texttt{listener}}$：

$$(\hat{x}, f_{\texttt{listener}}) = \begin{cases} (\texttt{noop}, 0) & P_{\max} < \theta_1 \\ (t, C(x)) & \theta_1 < P_{\max} < \theta_2 \\ (t, 0) & P_{\max} > \theta_2 \end{cases} \tag{2}$$

这种控制策略模仿了一位照料者：只有在理解语言学习者所指为何时才给予反馈，但会引导其生成质量更高的话语。当听者的信心很低时，它既不会选择目标图像，也不会提供语言输入，因为说话者的话语质量太低。与此同时，当听者的信心很高时，它会停止提供语言输入，因为此时认为没有必要进一步改进。

我们使用基于长短期记忆网络的模型来学习说话者话语的嵌入表示，并结合 ResNet 模型提取图像特征。由于我们希望听者至少能够模拟一位熟练掌握说话者试图习得之语言的使用者，因此在新环境中初始化听者之前，需要进行一些预训练。经过对数据集的初步实验后，设定控制器的值 θ_1，θ_2 以及 ResNet 参数。随后，使用小批量随机梯度下降训练听者的语言网络参数，以优化网络值

$$\theta_{\text{listener}} = \arg\max_{\theta} \mathbb{E}_{C\sim\mathcal{U}(I)^N} \frac{1}{N} \sum_{i=1}^{N} \log P_{\text{listener}}(i \mid U_i^*, C; \theta) \tag{3}$$

3 说话者训练过程

3.1 交际目标

我们将交际目标和从语言输入中学习建模为说话者网络的两个独立学习目标，并通过可调系数 λ 将二者结合起来。交际目标奖励 $\mathcal{O}_{CG}$ 只是训练回合中每场游戏的平均奖励（定义见 §2）。可能的输出话语空间是离散且不可微的，因此我们的说话者通过强化学习方法近端策略优化（Schulman 等，2017）学习其策略 π。

3.2 从语言输入中学习

我们将从语言输入中学习这一目标建模为说话者模型中听者输入的最大似然，并通过随机梯度下降优化这个连续函数。

$$\mathcal{O}_{LI} = \mathbb{E}_{x,C,u\sim\pi(u|x)} \log \pi(U^*_{f_{\text{listener}}(u,C)} \mid x, C) \tag{4}$$

然后，我们使用系数 $\lambda \in [0, 1]$ 将每项任务各自的学习目标结合起来。 :

$$\mathcal{O}_{\text{joint}} = \lambda\mathcal{O}_{CG} + (1-\lambda)\mathcal{O}_{LI} \tag{5}$$

接下来，我们介绍心智理论建模，以及一个额外的任务目标 $\mathcal{O}_{ToM}$。这是一个交叉熵目标，用于表示说话者的内部听者模型有多准确。与上述目标类似，我们将这些目标结合为整个说话者网络的单一学习目标。

4 介绍心智理论建模

我们在说话者架构中加入了一个基于 LSTM 的心智理论听者。该听者模型与说话者网络的其余部分一同训练，以学习实际听者的“心理状态”——即能够使其最大程度地复现听者为图像—话语对所分配概率的神经网络参数。它采用与实际听者相似的架构，将其通过 LSTM 网络学习到的句子嵌入与使用 ResNet 获得的预训练图像嵌入相结合。换言之，我们的心智理论听者旨在学习句子嵌入模型 θ_{ToM}，使得在给定话语的情况下，听者对其选择所分配的概率最大化。

$$\theta_{ToM} = \arg\max_{\theta} P_{ToM}(u_i|\hat{x}) \propto \exp(\theta^T_{ToM}(u_i) \cdot \texttt{ResNet}(\hat{x})) \tag{6}$$

我们通过让说话者采样候选话语，并借助心智理论听者对其重新排序，将心智理论引入说话者模型。我们的心智理论说话者首先从其分布中采样 N候选话语$U = \{u^{(i)}\}_{i=1}^{N}$。接下来，我们为每个话语生成说话者评分和听者评分:

$$P_{\texttt{speaker}}(u) \propto \prod_i P(u_i|u_1, u_2, \ldots u_{i-1}, x) \tag{7}$$

$$P_{\texttt{ToM}}(x|u) \propto \exp(\texttt{LSTM}^T(u)\texttt{Resnet}(x)) \tag{8}$$

这里，概率 $P(u_i|u_1,\ u_2,\ \ldots u_{i-1},\ x)$ 根据 1 计算。然后，我们使用听者权重超参数 w_l 合并这些评分，从而得到分配给每个话语的总评分。最后，我们选取评分最高的话语 u^b，即该评分的 argmax：

$$u^b = \arg\max_U (P_{\texttt{ToM}}(x|u^j)^{w_l} \cdot p_{\texttt{speaker}}(u^j)) \tag{9}$$

在实验中，我们使用 w_l 的三种不同设置训练模型。我们使用 $w_l = 0$ 训练模型，以将心智理论听者的影响与新模型架构的影响区分开来。我们使用 $w_l = 1$ 训练模型，使说话者和心智理论听者的输入权重相等。最后，我们训练的模型中，w_l 是任意设定的高常数 1000。在这种情况下，听者评分占主导地位，而我们的说话者实质上试图最大化 $P_{\texttt{listener}}(x|u^j)$，并以心智理论听者评分作为近似值。如果我们用外部听者的副本替换学得的心智理论听者，那么这一话语采样过程与 Andreas 和 Klein (2016) 中学得的理性言语行为（RSA）模型十分相近；该模型同样在指称游戏环境中训练语用说话者。我们将心智理论说话者与这些“RSA”说话者进行比较，以评估使用学得的近似模型而非实际听者所带来的影响。

最后，我们引入一个超参数 σ：说话者以概率σ 输出随机话语，以概率 $1-\sigma$输出 u^b。 σ 设置为随时间线性衰减；这种随机化旨在促进早期探索。

默认情况下，我们从一个未经训练的听者开始，并随着时间推移对其进行训练，使其模拟实际听者的输出。为了训练心智理论听者，除了$\mathcal{O}_{LI}$和 $\mathcal{O}_{CG}$之外，我们还引入第三个训练目标，即心智理论目标 $\mathcal{O}_{ToM}$，其定义为心智理论听者的分布与真实听者的分布之间的交叉熵损失。因此，我们训练它根据说话者的话语，对听者的选择 $\hat{x}$ 赋予更高的概率。形式化地说，

$$\mathcal{O}_{ToM} = \begin{cases} -\log P_{ToM}(\hat{x}|u) & P_{\max} > \theta_1 \\ 0 & P_{\max} < \theta_1 \end{cases} \tag{10}$$

请注意，这要求听者作出选择，因此，当实际听者未选择图像时，我们将该情况屏蔽，并将其设为零。我们使用相同的参数 θ_1和听者置信度 $P_{\max}$ 来实现这一点；这两个参数曾用于控制第 2 节中的语言输入。我们将其加入此前在第 5 节定义的联合学习目标，从而计算整个心智理论说话者系统的组合目标：

$$\mathcal{O}_{\text{joint}} = \lambda\mathcal{O}_{CG} + (1-\lambda)\mathcal{O}_{LI} + \mathcal{O}_{ToM} \tag{11}$$

然而，未经训练的听者可能会给话语生成过程带来显著噪声。为此，我们通过在固定的步数内将 w_l 从 0 线性增加到最终的 w_l 值，逐步退火听者的影响。这样一来，只有当心智理论听者至少具备一定能力时，说话者才会开始使用它。

5 提高干扰项难度

此前关于语言习得的研究发现，婴儿听到用于指代不同物体的语义相关标签时，起初会注视熟悉的物体；但随着时间推移，他们会学习更复杂的物体语义表征，从而适应更困难的任务（Bergelson 和 Aslin，2017a）。此外，Yuksekgonul 等人（2022）发现，在更难的干扰项上进行对比训练，能够提升视觉语言模型在涉及组合性的任务中的表现。我们假设，使用相似度更高的干扰图像，也可能迫使说话者生成更复杂的话语，因为它需要进一步区分不同图像。因此，我们在训练过程中生成了更相似的干扰图像，以期达到这一效果。具体而言，我们根据图像的视觉相似度和语义相似度计算图像之间的相似度排名。随后，在训练过程中选定一张“目标”图像后，我们抽取与目标图像相似度较高的图像，在训练时将其用作干扰项。

我们尝试了三种计算图像/标题相似度的方法：



Fig图 2：一个图像-标题对，以及按各种度量指标相似度排序的干扰项 s. Cap使用 TF-IDF 加权的 RoBERTa 嵌入计算文本和混合相似度。

• **视觉余弦相似度**。 我们根据预训练 ResNet 模型生成的图像嵌入的余弦相似度，计算视觉上最相似的图像。对于每张图像，我们保存与其视觉上最相似的后续 1000 张图像的相似度排名，并在训练过程中每当选中原始图像作为目标时，就从该集合中选取干扰项。

• **文本余弦相似度**。我们通过计算文本向量表示之间的余弦相似度来衡量标题的文本相似度。我们进行了两种方法的实验：（1）使用稠密向量，即通过 sentence-transformers 库中的预训练 RoBERTa 或预训练 CLIP 均值池化模型计算嵌入（Reimers 和 Gurevych，2019）；（2）使用稀疏向量，即为每个词构建独热向量。我们还尝试使用词频-逆文档频率（TF-IDF）对标题中每个词元对应的向量加权，以此计算标题相似度，从而提高较罕见实义词的权重。

• **视觉+文本相似度**。在同时使用图像和标题相似度的情况下，我们加入了描述权重参数 $w_c \in [0, 1]$，并用它计算图像相似度和标题相似度的加权平均。我们希望混合方法（$w_c = 0.5$）能够捕捉到在视觉和语义上都相似的干扰项。

我们还使用随机选取的或“简单”的干扰项训练模型，以此建立基线来研究干扰项难度的影响。图 2 展示了部分干扰项设置的示例。

6 实验设置与结果

6.1 实验设置

我们重点关注两项主要结果：说话者运用语言正确区分目标图像与干扰图像的能力，以及说话者所用语言的质量。

为评估前者，我们主要考察平均准确率，即听者选择正确指称对象的频率。我们通过流畅度、平均长度、显著词性的 F1 分数以及描述质量分数来评估说话者习得语言的质量。我们采用一种流畅度分数，用于衡量输出话语的语法质量（Kann 等人，2018）。该分数定义为从一元语法模型切换到预训练语言模型时对数概率的平均增益。

$$\text{fluency} = \frac{1}{|u|}(\ln(p_M(u)) - \ln(p_U(u))) \tag{12}$$

我们在训练集上训练一元语法模型 p_U，并对 GPT-2 大型版（Radford 等人，2019）p_M 进行微调。

除了流畅度和准确率外，我们还考察另外两项性能指标。为了更好地评估系统整体的描述质量，我们使用 Bird 等人（2009）提供的实现来计算 BLEU 分数（Papineni 等人，2002）。此外，为了评估习得的听者模型的准确率，我们计算心智理论准确率，其定义为：心智理论听者认为最有可能被选中的候选图像，实际是否为听者所选。

我们还会考察词性统计，以评估话语质量。我们使用 spaCy（Honnibal 等人，2020）为说话者话语和真实描述中的词语标注词性。我们用



基础说话者： 一名棒球运动员手里拿着一把雨伞。
在困难干扰项上训练的说话者： 一名棒球运动员在球场上拿着一根球棒。
在困难干扰项上训练的心智理论说话者： 一群男子正在户外打棒球。

图3：各种说话者模型生成的输出话语示例。此处所述的"困难干扰项"是利用 CLIP嵌入计算的混合指标选出的。使用更困难的干扰项进行训练以及结合心智理论进行训练，均能产生更准确、更流畅的话语。

表 1：各种心智理论说话者的模型性能和语言特征。

模型		性能				词性 F1			
心智理论权重	干扰项	Acc	BLEU	流畅度	ToM	形容词	介词	NOUN	VERB
基线（无心智理论）	Easy	0.81	0.20	1.50	N/A	0.16	0.52	0.41	0.38
基线（无心智理论）	Hard	0.81	0.24	1.87	N/A	0.24	0.58	0.46	0.45
黄金标准	N/A	0.92	1.00	2.52	N/A	1.00	1.00	1.00	1.00
Zero	Hard	0.83	0.26	1.99	0.81	0.22	0.64	0.49	0.47
正常	Hard	0.85	0.26	2.25	0.88	0.22	0.65	0.52	0.49
High	**Hard**	**0.88**	**0.27**	**2.23**	**0.89**	**0.22**	**0.66**	**0.52**	**0.50**
高 RSA	Hard	0.87	0.28	2.26	0.93	0.23	0.65	0.50	0.49
Zero	Easy	0.85	0.25	1.73	0.85	0.21	0.57	0.48	0.49
正常	Easy	0.88	0.26	2.09	0.91	0.21	0.64	0.50	0.52
High	**Easy**	**0.88**	**0.27**	**2.07**	**0.91**	**0.22**	**0.65**	**0.51**	**0.50**
高 RSA	Easy	0.89	0.29	1.91	0.94	0.17	0.65	0.52	0.49

这些标注，分别计算说话者话语与真实描述在各词性上的 F1 分数。借此，我们可以评估说话者能否有效识别目标图像中人工描述者认为显著的指称对象。某一词性上的 F1 分数越高，说明说话者越有能力识别属于该词性的概念。

图像和标题取自 Lin 等（2014）介绍的 MS COCO 数据集。MS COCO 2017 数据集提供的训练集-验证集-测试集划分扩展到了我们的实验设置中。

6.2 心智理论的影响

我们训练了配备心智理论听者的模型，分别采用正常权重（即在对输出话语进行重排序时，对说话者评分和听者评分赋予相同权重）和高权重（将听者权重任意设为 1000，使说话者实际上针对 $P(x|u^i)$ 进行优化）。为了将使用心智理论听者的影响与修改话语选择流程的影响区分开来，我们还训练了一个配备心智理论听者、但不给予其任何权重的模型。此外，我们还训练了一个不含任何心智理论组件的模型以供比较。对于简单和困难干扰项，我们都进行了上述训练；其中，困难干扰项是根据第 5 节所述的视觉特征与语义（CLIP）特征之间的混合相似度选出的。最后，我们将测试集中的真实描述提供给听者，以计算准确率和流畅度的金标准指标。

我们在表 1 中发现，当说话者模型仅根据心智理论听者评分对话语进行重排序时，性能有显著提升。这类"高权重心智理论"说话者模型在面对简单和困难干扰项时，准确率分别提升了 3.0% 和 4.6%。这表明，在说话者训练过程中加入影响力足够大的心智理论重排序器，能够提升说话者的表现，尽管在简单干扰项上训练时，相对提升幅度似乎要大得多。

表2：在各种干扰项上训练的说话者的性能和语言特征。我们仅展示图像描述和 Hybrid 相似度中性能最佳的变体，其余变体见 A.1。

模型 干扰项	性能 Acc	流畅度	词性 F1 ADJ	ADP	NOUN	VERB	平均值 长度
Base	0.81	1.50	0.16	0.52	0.41	0.38	8.97
黄金标准	0.92	2.52	1.00	1.00	1.00	1.00	10.79
图像	0.80	2.09	0.19	0.61	0.45	0.49	10.33
图像描述	0.86	2.01	0.19	0.62	0.49	0.48	10.04
混合	0.85	2.19	0.20	0.62	0.49	0.47	10.18

然而，我们发现，使用说话者心智理论分数对话语进行重排序的说话者模型，通常无法胜过训练时未使用心智理论听者的模型。

我们还发现，在简单干扰项（流畅度分数相对提高 15.6%）和困难干扰项（提高 11.6%）的训练中，使用高权重的心智理论听者会显著提升流畅度。使用正常或高权重的心智理论听者时，我们还观察到话语更长、结构也更复杂。此外，根据 BLEU 分数衡量，我们发现基线模型与高权重模型在整体图像描述能力方面的提升有限。不过，这些影响较为微妙，并不总能显著提高准确率，这表明心智理论准确率提升的主要驱动力是语用能力的增强。我们得出结论，在训练过程中使用影响力很强的心智理论听者，能够显著提升模型性能和流畅度。我们还可以通过定性观察看到心智理论带来的模型性能提升。如图 3 所示，在一个具有代表性的示例中，我们的心智理论说话者能够在流畅的话语中指出两个明显区分目标图像与干扰项的要素（即有多个男人在打棒球）。

最后，我们发现，心智理论听者能够成功地近似外部听者。采用学习得到的听者的模型与采用预训练听者的RSA模型在准确率和流畅度方面表现相当。由于RSA模型代表了说话者的听者模型所能达到的性能上限，这表明我们学习得到的听者对说话者非常有益。报告的较高心智理论准确率也体现了这一点，尤其是在表现最佳的模型中，即听者权重较高的模型中。这些定性和定量结果为心智理论能够在模拟语言习得中发挥重要作用提供了计算证据，这与人们关于心智理论在人类语言习得中可能发挥关键作用的假设相似。

6.3 干扰项难度的影响

如表2所示，我们总体上发现，在更困难的干扰项上训练的模型，其语言质量得到了显著提升。最大的增益体现在流畅度分数上，其中困难干扰项带来的提升幅度介于25%至46%之间。我们还发现，在更困难的干扰项上训练的模型所使用的词汇与真实描述更加相似，这一点是通过真实描述与所生成话语之间的F1分数来衡量的。在使用更困难干扰项训练的模型中，附置词、名词和动词的F1分数均显著更高。最后，与基准说话者相比，所有在困难干扰项上训练的说话者都会生成更加复杂的话语，其话语平均至少长一个词。这支持了我们的假设：当面临更大的环境压力时，说话者会通过使其语言更加精确、流畅和复杂来进行适应。图3中的一个代表性示例从定性角度展示了这一点：在困难干扰项上训练的说话者能够生成更加流畅、更加准确描述图像的话语（在该示例中，它正确地将图像中的物体识别为棒球棒，而不是雨伞）。

我们发现，使用不同类型的困难干扰项训练的模型，其语言差异较小。使用视觉相似干扰项训练的说话者模型流畅度最高，达到 2.094，且生成的话语最长。它们的动词选择也更准确，F1 分数更高。使用混合相似度或 Caption 相似度筛选出的干扰项训练的说话者，名词 F1 分数高达 0.49，而基础说话者模型为 0.41，这表明训练中使用语义相似的干扰项可能更有利于识别显著名词。我们发现，在较容易的干扰项上进行评估时，使用更困难的干扰项训练并不能持续提升模型性能。这表明语言的流畅度与其在

图像指称游戏环境中的适用性之间可能存在脱节。不过，使用语义相似度更高的干扰项训练的说话者，其准确率仍比基础说话者高出最多 5 个百分点，这表明使用某些较难的干扰项进行训练有助于提升性能。

7 相关工作

人类语言习得中的相似之处。 通过反复接触指称对象来学习语言，是心理学界广受欢迎的一种模型。Smith & Yu (2008) 发现，婴儿通过对词语—场景配对进行统计学习，能够消除判断场景中的哪个指称对象对应某个词语时的不确定性。Yu & Ballard (2007) 将社会因素纳入模型，考察了诸如注意力解读等“社会认知”能力，并将其视为心智理论的一种形式。Yu & Smith (2012) 使用头戴式摄像机研究照料者对早期语言习得的社会影响，发现当某个指称对象在视觉上占据主导地位时，照护者的反馈会直接促使婴儿习得语言。Bergelson & Aslin (2017b) 发现，相较于语义相似的图像，婴儿更容易区分彼此无关的图像。

语用建模。 Andreas & Klein (2016) 使用语用模型联合训练神经说话者与听者模型，使其能够参与一项指称游戏，在游戏中描述彼此对比的场景。他们还在这一设置中使用采样与重排序方法来选择话语。不过，他们使用的是真实听者，而非经学习得到的听者，来对话语重新排序。此外，我们将这一过程应用于语言习得计算模型。Nematzadeh et al. (2018) 创建了一个数据集，用于评估问答模型追踪不一致世界观的能力。他们发现，最先进的神经模型缺乏这种能力，说明它们无法解决涉及心智理论的任务。Monroe et al. (2017) 研究了视觉相似干扰项对颜色识别语用模型的影响，发现语用模型在最困难的设置下取得的提升最大。Vedantam et al. (2017) 也考察了在图像描述模型中加入语用组件的效果，具体而言，是加入一个用于更好地区分图像的内部模块。

涌现语言。 Lazaridou & Baroni (2020) 调查了多智能体通信中涌现语言的最新进展，并指出，进一步的发展有助于深度网络变得更具交互性和可解释性。Lazaridou et al. (2020) 将一个经过通用训练的语言模型置于具有任务特定奖励的多智能体环境中。与我们的工作类似，这使模型成为任务条件语言模型，作者认为它能够在视觉任务中更好地与人类交流。Chaabouni et al. (2022) 通过改变干扰项的数量，分析了任务难度对涌现通信任务的影响；结果表明，这会对评估期间的模型性能产生负面影响。Mu & Goodman (2021) 试图通过迫使说话者围绕代表抽象视觉概念的物体集合进行交流，提高指称游戏中所学语言的可解释性，并分析由此产生的涌现语言的组合性。

8 结论与未来工作

本文扩展了一个现有的计算框架，该框架用于模拟图像指称游戏环境中的语言习得过程。最值得注意的是，我们在说话者模型中加入了心智理论组件，使说话者能够借助经过学习的内部听者，以语用方式重新排列候选话语的顺序。我们还通过提高语义和视觉上更相似的干扰图像的权重，尝试增加干扰项难度。我们发现，将心智理论融入说话者模型能够提升说话者表现。我们还发现，加入难度更高的干扰项会促使语言变得更加复杂、流畅。

未来的研究可以衡量人类学习者与我们的模型在学习过程上的相似性，并考察本文实施的改动是否会带来更接近人类的学习过程。进一步的研究还可以探讨更动态的难度调整或课程设计所带来的影响——例如，研究在多种干扰项难度下训练的模型，能否调整其话语以适应具体语境。最后，我们可以通过改变听者架构，或考虑更具挑战性的场景（如物体指称游戏），在更复杂的环境中研究这些影响。我们鼓励机器学习和心理建模领域的研究者考虑在计算语言习得模型中进一步融入心智理论，这有助于开发更具语用意识的模型。

参考文献

Jacob Andreas 和 Dan Klein。借助神经网络听者与说话者进行语用推理。*2016*年自然语言处理实证方法会议论文集，1：1173–1182，2016年。

Nolan Bard、Jakob N Foerster、Sarath Chandar、Neil Burch、Marc Lanctot、H Francis Song、Emilio Parisotto、Vincent Dumoulin、Subhodeep Moitra、EdwardHughes 等。Hanabi挑战：人工智能研究的新前沿。《人工智能》，280:103216，2020年。

Elika Bergelson 和 Richard Aslin。一岁幼儿词语理解的语义特异性。《语言学习与发展》，13:1–21，2017年6月a。doi：10.1080/15475441.2017. 1324308。

Elika Bergelson 和 Richard N. Aslin。6个月婴儿词汇的性质与起源。《美国国家科学院院刊》，114(49):12916–12921，2017b。doi：10. 1073/pnas.1712966114。URL https://www.pnas.org/doi/abs/10.1073/pnas. 1712966114。

Steven Bird、Ewan Klein 和 Edward Loper。《Python自然语言处理：使用自然语言工具包分析文本》。O'Reilly，北京，2009。ISBN 978-0-596-51649-9。doi: http://my.safaribooksonline.com/9780596516499。URL http://www.nltk.org/book。

KrisCao、AngelikiLazaridou、MarcLanctot、Joel Z Leibo、Karl Tuyls 和 Stephen Clark。通过协商实现涌现式通信。《arXiv 预印本 arXiv:1804.03980》，2018。

Rahma Chaabouni、Florian Strub、Florent Altch´e、Eugene Tarassov、Corentin Tallec、Elnaz Davoodi、Kory Wallace Mathewson、Olivier Tieleman、Angeliki Lazaridou 和 Bilal Piot。大规模涌现式通信。载于《国际学习表征会议》，2022。URL https://openreview.net/forum?id=AUGBfDIV9rL。

何恺明、XiangyuZhang、任少卿和孙剑。用于图像识别的深度残差学习。发表于*IEEE*计算机视觉与模式识别会议论文集，第770–778页，2016年。

Matthew Honnibal、Ines Montani、Sofie Van Landeghem和Adriane Boyd。spaCy：Python中的工业级自然语言处理。2020年。doi：10.5281/zenodo.1212303。

Katharina Kann、Sascha Rothe和Katja Filippova。句子级流畅度评估：参考答案有所帮助，但并非必需！发表于第*22*届计算自然语言学习会议论文集，第313–323页，2018年。

Angeliki Lazaridou和Marco Baroni。深度学习时代涌现式多智能体通信。*CoRR*，abs/2006.02419，2020年。网址：https://arxiv.org/abs/2006.02419。

Angeliki Lazaridou、Alexander Peysakhovich 和 Marco Baroni。多智能体合作与（自然）语言的涌现。*arXiv* 预印本 *arXiv:1612.07182*，2016。

Angeliki Lazaridou、Anna Potapenko 和 Olivier Tieleman。多智能体通信与自然语言相遇：功能性语言学习与结构性语言学习之间的协同作用。*CoRR*，abs/2005.07064，2020。网址：https://arxiv.org/abs/2005.07064。

Tsung-Yi Lin、Michael Maire、Serge Belongie、James Hays、Pietro Perona、Deva Ramanan、Piotr Doll´ar 和 C Lawrence Zitnick。Microsoft COCO：情境中的常见物体。载于欧洲计算机视觉会议，第 740–755 页。Springer，2014。

Yinhan Liu、Myle Ott、Naman Goyal、Jingfei Du、Mandar Joshi、Danqi Chen、Omer Levy、Mike Lewis、Luke Zettlemoyer 和 Veselin Stoyanov。Ro{bert}a：一种稳健优化的 {bert} 预训练方法，2020。网址：https://openreview.net/forum?id=SyxS0T4tvS。

Ryan Lowe、Abhinav Gupta、Jakob Foerster、Douwe Kiela 和 Joelle Pineau。涌现式通信中监督与自我博弈之间的交互。在 国际学习表征会议，2019年。

Will Monroe、Robert X.D. Hawkins、Noah D. Goodman 和 Christopher Potts。语境中的颜色：用于具身语言理解的语用神经模型。计算语言学协会汇刊，5:325–338，2017年。doi：10.1162/tacl a 00064。URL https: //aclanthology.org/Q17-1023。Jesse Mu 和 Noah Goodman。泛化的涌现通信。在 M. Ran-zato、A. Beygelzimer、Y. Dauphin、P.S. Liang 和 J. Wortman Vaughan（编），神经信息处理系统进展，第34卷，第17994–18007页。Curran Asso-ciates, Inc.，2021年。URL https://proceedings.neurips.cc/paper/2021/file/9597353e41e6957b5e7aa79214fcb256-Paper.pdf。Aida Nematzadeh、Kaylee Burns、Erin Grant、Alison Gopnik 和 Tom Griffiths。评估问答中的心智理论。在 *2018*年自然语言处理实证方法会议论文集，第2392–2400页，比利时布鲁塞尔，2018年10月至11月。计算语言学协会。doi：10.18653/v1/D18-1261。URL https://aclanthology.org/D18-1261。Kishore Papineni、Salim Roukos、Todd Ward 和 Wei-Jing Zhu。BLEU：一种用于机器翻译自动评估的方法。在 计算语言学协会第*40*届年会论文集，ACL '02，第311–318页，美国，2002年。计算语言学协会。doi：10.3115/1073083.1073135。URL https://doi.org/10.3115/ 1073083.1073135。David Premack 和 Guy Woodruff。黑猩猩是否具有心智理论？行为与脑科学，1(4):515–526，1978年。Neil Rabinowitz、Frank Perbet、Francis Song、Chiyuan Zhang、SM Ali Eslami 和 Matthew Botvinick。机器心智理论。在 国际机器学习会议，第4218–4227页。PMLR，2018年。Alec Radford、Jeffrey Wu、Rewon Child、David Luan、Dario Amodei、Ilya Sutskever 等。语言模型是无监督多任务学习器。*OpenAI*博客，1(8):9，2019年。Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agar-wal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark、Gretchen Krueger 和 Ilya Sutskever。从自然语言监督中学习可迁移视觉模型，2021年。URL https://arxiv.org/abs/2103.00020。Nils Reimers 和 Iryna Gurevych。Sentence-BERT：使用孪生BERT网络的句子嵌入。在 *2019*年自然语言处理实证方法会议论文集。计算语言学协会，2019年11月。URL https://arxiv. org/abs/1908.10084。John Schulman、Filip Wolski、Prafulla Dhariwal、Alec Radford 和 Oleg Klimov。近端策略优化算法，2017年。URL https://arxiv.org/abs/1707.06347。Linda Smith 和 Chen Yu。婴儿通过跨情境统计快速学习词语-指称映射。认知，106(3):1558–1568，2008年。ISSN 0010-0277。doi：https://doi.org/ 10.1016/j.cognition.2007.06.010。URL https://www.sciencedirect.com/science/ article/pii/S0010027707001795。Michael Tomasello。构建语言：基于使用的语言习得理论，第1卷。哈佛大学出版社，2005年。Ramakrishna Vedantam、Samy Bengio、Kevin Murphy、Devi Parikh 和 Gal Chechik。来自上下文无关监督的上下文感知图像描述。*CoRR*，abs/1701.02870，2017年。URL http: //arxiv.org/abs/1701.02870。Chen Yu 和 Dana H. Ballard。早期词汇学习的统一模型：整合统计线索与社会线索。神经计算，70(13):2149–2165，2007年。ISSN 0925-2312。doi：https://doi.org/ 10.1016/j.neucom.2006.01.034。URL https://www.sciencedirect.com/science/ article/pii/S092523120600508X。第3届发展与学习国际会议（ICDL 2004）精选论文 时间序列预测竞赛：CATS基准。

Chen Yu 和 Linda B. Smith. 幼儿的具身注意力与词汇学习。认知，125(2):244–262, 2012。ISSN 0010-0277。doi: https://doi.org/10.1016/j.cognition. 2012.06.016。URL https://www.sciencedirect.com/science/article/pii/ S0010027712001369。

Mert Yuksekgonul、Federico Bianchi、Pratyusha Kalluri、Dan Jurafsky 和 James Zou。视觉语言模型何时以及为何会表现得像词袋，以及该如何应对？2022。URL https://arxiv.org/abs/2210.01936。

Hao Zhu、Graham Neubig 和 Yonatan Bisk。通过建模心智理论实现小样本语言协调。发表于国际机器学习会议，第 12901–12911 页。PMLR，2021。

Hao Zhu、Yonatan Bisk 和 Graham Neubig。通过交际目标和语言输入模拟语言学习。认知科学学会年会论文集，44：1351–1358，2022。

A 附录

A.1 干扰项难度的完整影响

在表2中，我们选取了参与训练表现最佳模型的图像描述和混合干扰项。正如承诺的那样，我们在此报告涵盖更广泛干扰项变体的实验结果。

表3：使用所有干扰项变体训练的说话者的表现和语言特征。

模型 干扰项	性能 Acc	 流畅度	词性 F1 ADJ	 ADP	 NOUN	 VERB	平均值 长度
Base	0.81	1.50	0.16	0.52	0.41	0.38	8.97
黄金标准	0.92	2.52	1.00	1.00	1.00	1.00	10.79
图像	0.80	2.09	0.19	0.61	0.45	0.49	10.33
CLIP 图像描述	0.83	1.87	0.18	0.57	0.46	0.43	9.56
RoBERTa 图像描述	0.83	1.93	0.20	0.60	0.48	0.47	10.06
RoBERTa 图像描述 TFIDF	0.86	2.01	0.19	0.62	0.49	0.48	10.04
图像描述 TFIDF	0.83	2.05	0.21	0.63	0.46	0.45	10.30
混合 CLIP	0.81	1.87	0.24	0.58	0.46	0.45	9.78
混合 RoBERTa	0.83	1.94	0.20	0.61	0.48	0.47	10.09
混合 RoBERTa TFIDF	0.84	2.09	0.20	0.65	0.48	0.46	10.29
混合 TFIDF	0.85	2.19	0.20	0.62	0.49	0.47	10.18